PCT/AU2004/000857

Patent Office
Canberra

REC'D 28 JUL 2004
WIPO PCT

I, JULIE BILLINGSLEY, TEAM LEADER EXAMINATION SUPPORT AND SALES hereby certify that annexed is a true copy of the Provisional specification in connection with Application No. 2003903306 for a patent by COMMON GROUND PUBLISHING PTY LTD as filed on 27 June 2003.

COMMONWEALTH OF AUSTRALIA
AUSTRALIA
PATENT OFFICE

WITNESS my hand this
Fifteenth day of July 2004

J. Billingsley

JULIE BILLINGSLEY
<u>TEAM LEADER EXAMINATION</u>
<u>SUPPORT AND SALES</u>

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

P/00/009 25.5.91
Regulation 3.2

AUSTRALIA

Patents Act 1990

PROVISIONAL SPECIFICATION

Invention Title: Method and apparatus for extending the range of useability of ontology driven systems and for creating interoperability between different mark-up schemas for the creation, location and formatting of digital content.

The invention is described in the following statement:

Method and apparatus for extending the range of useability of ontology driven systems and for creating interoperability between different mark-up schemas for the creation, location and formatting of digital content.

FIELD OF THE INVENTION

The present invention pertains to a method and an apparatus for extending the range of useability of ontology driven systems beyond the scope of its original design, and for creating interoperability between different mark-up schemas for the location and formatting of digital content.

TECHNICAL FIELD

A key background is the emergence of the Internet and the World Wide Web as widespread information and communications technology in the mid 1990s, and more recently in 1998, the invention of XML or Extensible Mark-up Language. Both the web and XML are derivatives of a much older computer technology, Standardised General Mark-up Language (SGML), which originated in the IBM laboratories in the early 1970s as a framework for the documentation of technical text, such as computer manuals. In the late 1980s, Tim Berners-Lee began working on a vastly simplified version of SGML, which was to become the heart of the World Wide Web: HyperText Mark-up Language, or HTML. The key deficiency of HTML (progressively removed by the current version 4 of HTML) was that it was somewhat of a conceptual jumble, mixing historical typesetting tags (presentational concepts) with structural and semantic tags.

XML however, made an enormous conceptual leap, in two regards. First, it is not a mark-up language, but a mark-up language for mark-up languages—a place where, in other words, any mark-up language could be created. Second, XML rigorously separates mark-up for structure and semantics from presentation—which occurs independently in a 'stylesheet transformation' area. The great benefit of the XML approach is that content can be 'multi-purposed' by means of different stylesheet

transformations. A page of text, for example, can be rendered as a web page, or a printed page, or as an image on a portable reading device, or as synthesised voice.

XML has since become ubiquitous. To take the example of the publishing industry, a number of key XML-based standards have emerged. These are in the areas of:

a) data entry and typesetting (Unicode and DocBook);
b) print rendering (Job Definition Format);
c) electronic rendering (XHTML, Open eBook, Digital Talking Book); B-2-B ecommerce for publishers and booksellers (the Online Information Exchange standard);
d) Library cataloguing (principally the Library of Congress MARC, MODS and METS standards);
e) Digital Rights Management (Extensible Digital Rights Management Language and the new MPEG21 Rights Data Dictionary); and
f) e-learning (the Shareable Content Object Reference Model and the Instructional Management Systems standards).

These standards are an example of what is now called the 'semantic web'. Each XML schema is 'ontology' consisting of a content tagging schema which describes the scope of a particular software application. These are the basis either of Document Type Definitions (or DTDs in XML file format) or database structures (which can, in turn produce exports into XML files based on the database structure). Tim Berners-Lee predicts that this is the next great step in the development of the internet, and one which promises more accurate resource discovery, machine translation and eventually, artificial intelligence.

There is one great barrier to this vision, and that is the problem of interoperability. Even though each standard or XML DTD has its own functional purpose, there is a remarkable amount of overlap between these standards. The overlap, however, often involves the use of tags in mutually incompatible ways. Our extensive preliminary mapping of the seventeen major standards that apply in just one industry—the publishing industry—shows that, on average, each standard shares seventy per cent of its semantic range with neighbouring standards. Despite this, it is simply not

possible to transfer data from one standard to another as each standard has been designed as its own independent, stand alone DTD. This, in fact, points to one of the key deficiencies of XML as a meta-mark-up framework: it does not provide a way for DTDs to relate to each other. In fact, its very openness invites a proliferation of DTDs, and with this proliferation, the problem of interoperability compounds itself.

This produces a practical, commercial problem. In the book publishing and manufacturing supply chain, for instance, different links in the chain use different standards: typesetters, publishers, booksellers, printers, manufacturers of electronic rendering devices and librarians. This disrupts the digital file flow, hindering supply chain integration and the possibilities of automating key aspects of supply chain, manufacturing and distribution processes. Precisely the same practical problems of interoperability are now arising in other areas of the electronic commerce environment.

BACKGROUND ART

A known but not commonly known task in today's IT world involves sharing data between systems. XML has emerged as the so-called 'syntactic sugar' to facilitate this task. As an example, Company A may have a commercial obligation to provide Company B with metadata about a series of documents, such as their titles, authors, classificatory categories and ISBNs. Both parties must agree on a common DTD to allow this to happen, which may be devised by the parties or based on an existing standard. In addition, each party must map their internal systems to this common DTD. Finally a further set of information—security constraints, transactional characteristics, network protocols and messaging conditions (whether responses must be synchronous or asynchronous)—must be agreed to before the metadata can be transferred. This complexity arises in the relatively simple transfer of information between two conferring parties.

However, in a scenario where there are many more than two parties, where the information is not covered by a single standard, where the resources and skills of the parties cannot facilitate costly and time-consuming integration, a different approach is needed—one which caters for the complexity of the messages, while providing tools

which simplify the provision and extraction of metadata. This approach is one which has been termed semantic interoperability. It involves providing a systematic mapping of associated XML standards to a common XML 'mesh', which must track semantic overlays and gaps, schema versioning, namespace resolution, language and encoding variances, and which must provide a comprehensive set of rules covering the data transfer—such as security, transactional and messaging issues.

The idea of a 'meta-schema'—a schema to cover other related schemas—was initially considered to be sufficient. Research has demonstrated, however, that this is not enough, being subject to many of the same problems as the individual schemas being mapped—versioning, terminological differences and so on.

Mark-up ontologies or software tagging systems use a variety of encoding formats, including Extensible Mark-up Language (XML) and Resource Definition Framework (RDF). Ontologies promise to overcome two of the most serious limitations of the World Wide Web:

1. the fact that searching is simply for semantically undifferentiated strings of characters; and

2. the fact that rendering alternatives are mostly limited by data entry methods—printed web pages do not live up to the historical standards of design and readability of printed text, and alternative non-visual renderings, such as digital talking books are at best poor.

Specific ontologies are designed to provide more accurate search results than is the case with computer or web-based search engines. Examples include the Dublin Core Metadata Framework and MARC electronic library cataloguing system. However metadata harvested in one scheme can not be readily or effectively be used in another.

Specific ontologies are also designed for a particular rendering option. For instance, amongst ontologies describing the structure of textual content, HTML is designed for use in web browsers, DocBook for the production of printed books, Open eBook for rendering to hand held reading devices and Digital Talking Book for voice synthesis. Very limited interoperability is available between these different ontologies for the

structure of textual data, and only then if it has been designed into the ontology and its associated presentational stylesheets.

Furthermore, it is not practically possible to harvest accurate metadata from data, as data structuring ontologies and ontologies for metadata are mutually exclusive. The field of the semantic web attempts to improve the inherent deficiencies in current digital technologies both in the area of resource discovery (metadata-based search functions) and rendering (defining structure and semantics in order to be able to support, via stylesheet transformations, alternative rendering options).

Its success, however, has been very limited, primarily because of the semantic dissimilarities between overlapping ontologies and because of the limited rendering options catered for in ontologies which define data structure. At most, one-to-one, schema-to-schema 'crosswalks' have been created.

Creating a single crosswalk is a large and complex task. As a consequence, the sheer number of significant overlapping ontologies in a domain presents a barrier to achieving interoperability. For instance, our research has identified some 17 major ontologies pertaining to the domain of authorship and publishing. Using the 'crosswalk' approach, every tag in a schema needs to be mapped tag by tag against every tag in every other schema with which interoperability is required.

Each crosswalk in fact involves two translations: Ontology A defined tag by tag in terms of Ontology B, and Ontology B defined tag by tag in terms of Ontology A. Using the crosswalk method, the number of mappings to achieve interoperability between N tagging schemas is $2\{(N/2)(N-1)\}$. In a terrain encompassed 17 ontologies, 272 crosswalks would be required (see FIG. 1). Moreover, new ontologies are regularly emerging and each new ontology increases exponentially the scale of the task of achieving interoperability.

The present invention addresses fundamental problems that currently arise in the area of interoperability of data and metadata. These can be summarised as follows:
1. The failure of 'the semantic web' to improve on the search mechanisms of on computers and the Internet across even similar domains of knowledge, information

and data. As a consequence, searching still functions primarily on the basis of a semantically agnostic process of matching of strings of characters.

2. There is limited interoperability between ontologies for metadata tagging, and when there is, it is a consequence of the laborious manual crosswalks approach.

3. There is a limited range of rendering options, even when mark-up for structure and semantics are separated from the rendering apparatus of the stylesheet.

SUMMARY OF THE INVENTION

The invention involves a method and an apparatus for extending the use of extant ontology driven software and digital file mark-up schemas into overlapping domains, such as RDF or XML-instantiated ontologies. Our invention arises from the technical and commercial logistics of structuring and rendering text, digital resource discovery, library cataloguing, ecommerce, digital rights management and e-learning. However, the method and apparatus of our invention is applicable to any other contexts demanding interoperability of tagged data. In application, the apparatus creates functionalities for data framed within the paradigm of one schema which extend well beyond those originally conceived by that schema. The invention creates interoperability between schemas, allowing data originally designed for use in one schema for a particular set of purposes to be used in another schema for a different set of purposes.

In accordance with the invention there is provided a method and apparatus for extending the range of useability of ontology driven systems and for creating interoperability between different mark-up schemas for the creation, location and formatting of digital content, the method includes the steps of:

a) having a database or datafile of digital content in a Document Type Definition of the first digital mark-up or computer software ontology able to be outputted in a selected format allowed by the first digital mark-up or computer software ontology;

b) organising digital mark-up or computer software tags of the first digital mark-up or computer software ontology into an overarching interlanguage ontology capable of absorbing and incorporating at least one other digital mark-up or computer software ontology;

c) automatically translating a Document Type Definition of the first digital mark-up or computer software ontology into a translated interlanguage Document type Definition;

d) selecting one of the at least one other digital mark-up or computer software ontology;

e) automatically translating the translated interlanguage Document Type Definition into a Document Type Definition of the selected other digital mark-up or computer software ontology thereby allowing information in the database or datafile format to be outputted in the required selected format allowed by the selected other digital mark-up or computer software ontology.

The step of organising digital mark-up or computer software tags of the first digital mark-up or computer software ontology into an overarching interlanguage ontology capable of absorbing and incorporating at least one other digital mark-up or computer software ontology includes the steps of indexing according to the following rules:

(i) providing a first level of granularity such that tags which represent data at a finer level of delicacy in Ontology X produce automatically recomposed data in Ontology Y which manages the same data at a higher level of semantic aggregation.

(ii) providing a lowest common denominator semantics such that, when data has been data marked up with a pair of tags that can be interpreted to be closely synonymous but not identical, the narrower semantics of the two tags is operationalised.

(iii) providing contiguous domains wherein tags can be aggregated and aligned by virtue of the fact that they relate to semantically exclusive data....

(iv) providing subset schemas within a tag such that a whole new domain identified by within Ontology Q or within a defined area of ontology Q can be mapped within a single tag in Ontology R....

The Mark-up Language of the invention (CGML) is a unique kind of DTD. In fact, although it is technically a DTD, it is a DTD of a fundamentally different order to any other. It does not have an independent life as a DTD. Rather, it is a uniquely designed and constructed apparatus whose semantic life is derived solely from other DTDs and whose operational realisation is found within other DTDs. This adds

another fundamental layer to the bifurcation of DTDs representing structure and semantics and DTDs representing rendering or presentational alternatives (stylesheets). The interlanguage apparatus is a DTD which does not manage structure and semantics per se; rather it automatically manages the structure and semantics of structure and semantics. Its mechanism, in other words, is metastructural and metasemantic. We have named its underlying mechanism the 'interlanguage' apparatus. Although developed in the case of one particular instantiation of problem of interoperability—for the electronic standards that apply to the publishing supply chain—the core technology is applicable to the more general problem of interoperability characterised by the semantic web and electronic commerce.

By filtering standards that don't talk to each other through the 'interlanguage' mechanism for database and document tagging which forms a core part of the invention, the method now allows talk between unrelated schemas. This produces immediate supply chain efficiencies through the automated transition of digital content from one electronic standard to another. It also provides for the multi-purposing of digital content, so that data is fully interoperable across all the full range of functional uses possible in the digital production and transmission of content. Three such applications for this technology are publishing, conference and learning management software products. There are many others, well outside the domain of textual content.

The invention described here allows metadata newly created through its apparatus to be interpolated into any number of metadata schemas. It also provides a method and apparatus by means of which data harvested in one metadata schema can be imported into another.

This invention is a unique method providing a highly flexible rule-based system for automatically inter-connecting XML schemata, in a way that each term of a schema could be related to one or more terms of one or more other schemata, with a rule-driven mechanism determining the nature of the relation.

Other possibilities of this technology are in the areas in which the semantic web has so much—as yet unfulfilled—promise. This includes: indexing, cataloguing and metadata systems; product identification systems; systems for the production, manufacture and distribution of copyright digital content; knowledge and content management systems; systems for multi-channelling content providing for disability access, for instance; machine translation from one natural language to another; and artificial intelligence.

The method and apparatus can be used for separating underlying ontology within ontology-based software or mark-up schema from its domain-specific application.

In one form there is provided a method including the step of forming the expression of meaning function so that it automatically manufactures tangible, expressed meanings in different formats, such as to a web-browser rendering to a computer screen, a typesetting device rendering to print, hand-held reading devices, personal digital assistants and mobile phones rendering to portable screen, or as digital talking book rendering as synthesised speech.

The method can include the step of forming the definition of metadata so that it is automatically generated in different formats, for different purposes; and creates different uses of the digital or physical content to which that metadata refers—for instance, as a library cataloguing record, a learning object, a digital rights management record; or an ecommerce record.

The present invention further pertains to a method and a system which allows data that has been entered into a computer to be used in multiple ways, even if these ways were not intended at the point of data entry or inherent to the data entry method. These varied uses may involve alternative forms of data rendering and multiple forms of metadata representation. It also allows interoperability of data entered in one software or mark-up schema, with other software schemas, even if the semantic range and functions of the original schema are narrower than those for which the data, using this invention, is now used. The technical fields in which this invention operates are metadata and mark-up schemas in computer software systems. The invention

automates the rendering of digital content in multiple and alternative data and metadata frameworks, and recognisable by different software systems.

This method upon which this invention is based is an 'interlanguage' built into the functioning of a computer system. By virtue of the operation of interlanguage tags forming an intermediating ontology, a full set of and indefinitely extensible series of crosswalks with a domain can be achieved easily and effectively (see FIG. 2). The interlanguage automates the crosswalk-to-crosswalk process. If tag <x> in Ontology A translates into tag <q> in the interlanguage, and tag <y> in Ontology B also translates into tag <q> in the interlanguage, then an automated tag translation from Ontology A to Ontology B can be achieved. The practical effect of the interlanguage is to add the functionality of Ontology A to the functionality of Ontology B, even though interoperability of data and functionalities may not have been conceived by the designers of a particular ontology, nor anticipated by the users entering data within the framework of each ontology. By means of this invention, a practical mechanism is created by means of which 272 crosswalks can be replaced by seventeen crosswalks to the interlanguage (see FIG. 3).

This invention achieves the results for which it has been designed by means of the following two mechanisms:

1. For data already residing in XML or RDF schemas or ontologies, it automatically passes that data through a filter apparatus using the interlanguage mechanism, into other schemas and ontologies even through the data had not originally been designed for the destination schema. The filter apparatus is driven by semantic and syntactical mechanics, and throws up queries whenever an automated translation of data is not possible in terms of those semantic rules.
2. For new data, the filter apparatus provides full automation of interoperability as the semantic and syntactical rules built into the software code from which the apparatus is constructed.

The apparatus is able to read tags automatically, and thus interpret the data which has been marked up by these tags, according to two overarching mechanisms, and a

number of submechanisms. The two overarching mechanisms are the superordination mechanism and the composition mechanism.

The superordination mechanism automatically constructs tag-to-tag 'is a ...' relationships. Within the superordination mechanism, there are the submechanisms of hyponymy ('includes in its class ...'), hyperonymy ('is a class of ...'), co-hyperonomy ('is the same as ...'), antonymy ('is the converse of ...') and series ('is related by gradable opposition to ...').

The composition mechanism automatically constructs tag-to-tag 'has a ...' relationships. Within the composition mechanism, there are the submechanisms of meronymy ('is a part of ...'), co-meronymy ('is integrally related to but exclusive of ...'), consistency ('is made of ...'), collectivity ('consists of ...').

These mechanisms for data interpolation are illustrated in the lower half of FIG. 4. These mechanisms are fully automated in the case of new data formation within any schema, in which case, deprecation of some aspects of an interoperable schema may be automatically requested at the point of data entry.

In the case of legacy data generated in schemas without anticipation of, or application of, the interlanguage mechanism, data can be imported in a partially automated way. In this case, tag-by-tag or field-by-field queries are automatically generated according to the filter mechanisms of:

taxonomic distance (automatically testing whether the relationships of composition and superordination are too distant to be necessarily valid),

levels of delicacy (whether an aggregated data element needs to be disaggregated and re-tagged),

potential semantic incursion (identifiable sites of ambiguity), and

translation of silent into active tags or vice versa (at what level in the hierarchy of composition or superordination data needs to be entered to effect superordinate transformations).

This mechanism for data interpellation is illustrated in the upper half of FIG. 4.

In one scenario, a quantum of legacy source data is provided to the apparatus, marked up according to the schematic structure of a particular source ontology. The apparatus then reads the structure and semantics ontology immanent in the data, interpreting this both from DTD and the way the DTD is realised in that particular instance. It applies four filters: a delicacy filter, a synonomy filter, a contiguity filter and a subset filter. The apparatus is able to read into the DTD and its particular instantiation an inherent taxonomic or schematic structure. Some of this is automated, as the relationships of tags is unambiguous based on the readable structure of the DTD and evidence drawn from its instantiation in a concrete piece of data. The apparatus also be able know of points at which it is possible there might be ambiguity, in this case throw up a structured query to the user. Each human response to a structured query becomes part of the memory of the apparatus, with implications drawn from the user response and retained for later moments when interoperability is required by this or another user.

On this basis, the apparatus interpellates the source data into the interlanguage format, whilst at the same time automatically 'growing' the interlanguage itself based on knowledge acquired in the reading of the source data and source ontology.

Having migrated into the interlanguage format, the data is then reworked into the format of the destination ontology. It is rebuilt and validated according to the mechanisms of superordination (hyponymy, hyperonymy, co-hyperonomy, antonymy and series) and composition (meronymy, co-meronymy, consistency, collectivity). A part of this process is automated, according to the inherent structures readable into the destination ontology, or previous human readings that have become part of the accumulated memory of the interlanguage apparatus. Where the automation of the rebuilding process cannot be undertaken by the apparatus with assurance of validity (when a relation is not inherent to the destination DTD, nor can it be inferred from accumulated memory in which this ambiguity was queried previously), a structured query is put to the user, whose response in turn becomes a part of the memory of the apparatus, for future use.

On this basis, the data in question is interpolated into its destination format. From this point, it can be used in its destination context or DTD environment, notwithstanding the fact that the data had not been originally formatted for use in that environment.

In another scenario, new data might be constructed in a source ontology which has already become 'aware' by means of previous applications of the interlanguage mechanism as a consequence of the application of the apparatus described above. In this case, the mechanism commences with the automatic interpellation of data, as the work of reading and querying the source ontology has already been performed. In these circumstances, the source ontology in which the new data is constructed becomes a mere facade for the interlanguage, taking the form of a user interface behind which the processes of subordination and composition occur.

Key operational features of this invention are:

1. The capacity to absorb effectively and easily new ontologies which refer to domains of knowledge, information and data that substantially overlap (vertical ontology-over-ontology integration). The invention is capable of doing this without the exponential growth in the scale of the task characteristic of the existing crosswalk method.
2. The capacity to absorb ontologies representing new domains that do not overlap with the existing range of domains and ontologies representing these domains (horizontal ontology-beside-ontology integration).
3. The capacity to extend indefinitely into finely differentiated subdomains within the existing range of domains connected by the interlanguage, but not yet this finely differentiated (vertical ontology-within-ontology integration).

One embodiment of this invention is a publishing system by means of which creators and publishers enter metadata which is interoperable across an extensible range of metadata systems.

Another embodiment of this invention is a text editor which captures the structure and semantics of textual and other data in such a way that it is interoperable across an extensible range of rendering formats and media.

Another embodiment of this invention is an ontology building apparatus by means of which application- and use-specific semantics can be crafted which conform to the

underlying semantic apparatus, and which as a consequence guarantees interoperability and automates alternative metadata retrieval and rendering options.

Another embodiment of this invention is a multilingual and multi-script translation apparatus, by means of which ontologies and software systems originally conceived and mapped in one language can be applied in a way conforming to their original semantics in a language for which they were not designed.

Another embodiment of this invention is an apparatus for structural and semantic mark-up which adds accuracy to machine translation by providing markers intelligible to the translation as a controlled vocabulary, based on their origins and recognisable ontologies.

In one of a broad range of possible instantiations, the invention tackles one of the fundamental issues of the 'semantic web'—the problem of interoperability between overlapping and related electronic standards and particularly in the area of publishing—how to relate standards in the areas of 1) typesetting and content capture, 2) electronic rendering, 3) print rendering, 4) B-2-B ecommerce, 5) digital rights management, 6) e-learning, 7) internet resource discovery and 8) cataloguing. The underlying 'interlanguage' mechanism of the Mark-up Language can be seen to extend the useability of content across multiple standards. The method ameliorates the enormous problem of interoperability in general, not just in publishing but in other areas of the semantic web.

Section 1: Core Ontology-Building Tool

This section of the instantiation of the interlanguage apparatus for the publishing industry locates CGML in a core piece of ontology building software, CommonGroundLEXICOGRAPHER. This piece of software defines and determines:

- Database structures for storage of metadata and data.
- XML document inputs.
- Synonyms across the tagging schemas for each standard against which CGML maps.

- Two definitional layers for every tag: underlying semantics and application-specific semantics.
- Export options into an extensible range of electronic standards expressed as XML DTDs.

The essential operative feature of this section is to provide the core apparatus for managing the interlanguage mechanism that is at the heart of this invention. It manages the superordination and compositional mechanisms described earlier, as well as providing an interface for domain-specific applications in which interoperability is required (such as publishing or learning management systems).

Section 2: eCommerce Interoperability

This section builds and tests ecommerce functionalities by means of CGML, principally ONIX, or the Online Information Exchange standard, initiated in 1999 by the Association of American Publishers, and subsequently developed in association with the British publishing and bookselling associations. The purpose of ONIX is to capture data about a work in sufficient detail to be able automatically to upload new book data to online bookstores such as Amazon.com, and to communicate comprehensive information about the nature and availability of any work of textual content. This sits within the broader context of interoperability with ebXML, an initiative of the United Nations Centre for Trade Facilitation and Electronic Business.

Key areas of technical improvement in this section include:

- Creating data which exports automatically into the book production supply chain.
- Creating data which works within overarching ecommerce protocols.

The essential operative feature of this section is to create a fully interoperable mechanism for managing ecommerce transactions related to digital content.

Section 3: Interoperability of Cataloguing, Indexing and Resource Discovery

This section builds and tests interoperabilities for cataloguing, indexing and resource discovery within the CGML 'interlanguage' mechanism. The MARC (Machine

Readable Catalogue) format was initially developed in the 1960s by the US Library of Congress. Most recently, MARC has been translated into three XML formats: a full version; a cut-down version under the name MODS (the Metadata Object Description Schema; and a standard specifically for the identification, archiving and location of electronic content, METS: the Metadata Encoding and Transmission Standard. In similar territory, although taking somewhat different approaches to MARC, are Biblink and Encoded Archival Description Language. In the indexing and resource discovery areas, Dublin Core has gained wide international acceptance. Although there are some isolated and ad hoc standard-to-standard 'crosswalks', no generalised interoperability across these standards has been achieved, nor with other standards related to other functionalities around textual and other creative content.

Key areas of technical improvement include:

- Creating a system which creates valid records on the fly across variant cataloguing, indexing and resource Discovery frameworks.

The essential operative feature of this section is to create a fully interoperable mechanism for managing the indexing and cataloguing digital content.

Section 4: Tool for the Capture of Text as Structured Data, Interoperable with Print and Electronic Rendering Standards

This section builds and tests interoperabilities for capturing and rendering text within the CGML 'interlanguage' mechanism. A number of electronic standards have been created for the purpose of describing the structure of text in order to facilitate its rendering to alternative formats. Unicode is designed as a universal multilingual character encoding standards; HTML4 and XHTML are designed primarily for rendering transformations through web browsers; the OASIS/UNESCO sanctioned DocBook standard is for structuring text to be rendered electronically or to print; Open e-Book is for rendering to hand-held reading devices; and Digital Talking Book is for rendering to audio as synthesised speech. Although there are some specific interoperabilities built into particular standards, there is as yet no generalised interoperability across rendering standards.

Key areas of technical improvement include:

- Development of a workable author-friendly mark-up interface.
- Metadata automatically generated from the structural and semantic mark-up of the data.
- Multi-channelling of content into formats defined by variant standards: to print, to screen, to audio.

The essential operative feature of this section is to create a fully interoperable mechanism for managing the structural and semantic mark-up of digital content.

Section 5: Automated Workflow into Digital and Offset Print Manufacture.
This section builds and tests interoperabilities for print manufacture within the CGML 'interlanguage' mechanism. The Job Definition Format is rapidly becoming the universal standard for the printing industry, as a digital addendum to offset print, and as the driver of digital print. Interoperability of JDF with other standards mean, for instance, that a book order triggered through an online bookstore (the ONIX space) generate a JDF wrapper around a content file as an automated instruction to print and dispatch a single copy.

Key areas of technical improvement include:

- Developing automated cross standards and cross supply chain manufacturing mechanism.

The essential operative feature of this section is to create a fully interoperable mechanism for managing printing of digital content.

Section 6: e-Learning Interoperability Mechanism
This section builds and tests interoperabilities for elearning environments within the CGML 'interlanguage' mechanism. Cutting across a number of areas—particularly rendering and resource discovery—are tagging schemas designed specifically for educational purposes. EdNA and the UK National Curriculum Metadata Standard are

both variants of Dublin Core. Rapidly rising to broader international acceptance, however, is the Instructional Management Systems Standard and the related Shareable Content Object Reference Model. Not only do these standards specify metadata to assist in resource discovery. They also build and record conversations around interactive learning, manage automated assessment tasks, track learner progress and maintain administrative systems for teachers and learners. The genesis of IMS was in the area of metadata and resource discovery, and not the structure of learning texts. One of the pioneers in the area of structuring and rendering learning content (building textual information architectures specific to learning and rendering these through stylesheet transformations for web browsers) was Educational Modelling Language. More recently, EML has been grafted into the IMS suite of schemas and renamed the IMS Learning Design Specification. The e-learning components of CGML we have name Learning Design Language—which crosses all e-learning standards.

There are levels of technical improvement achieved, particularly:

- Achieving functional interoperability across e-learning standards;
- Integrating e-learning standards with broader resource discovery, rendering, ecommerce, digital rights and other standards.

The essential operative feature of this section is to create a fully interoperable mechanism for integrating digital content into learning management systems.

Section 7: Achieving Digital Rights Interoperability

This section builds and test interoperabilities for digital rights management within the CGML 'interlanguage' mechanism. Digital Rights Management involves the identification of copyright owners and legal purchasers of creative content; it can also involve systems of encryption by means of which content is only accessible to legitimate purchasers; and systems by means of which content can be decomposed into fragments and recomposed by readers to suit their specific needs. The <indecs>, or Interoperability of Data in E-Commerce Systems framework was first published in 2000, the result of a two year project by the European Union to develop a framework for the electronic exchange of intellectual property (<indecs> 2000). The conceptual basis of <indecs> has more recently been applied in the development of the Rights

Data Dictionary for the Moving Pictures Expert Group's MPEG-21 framework for distribution of electronic content. From these developments and discussions, a comprehensive framework is now emerging, capable of providing mark-up tools for all manner of electronic content. Amongst the other tagging schemas marking up digital rights, Open Digital Rights Language is an Australian initiative which has gained wide international acceptance and acknowledgement. And XrML, or Extensible Rights Mark-up Language was created in Xerox's PARC laboratories in Paulo Alto. Its particular strengths are in the areas of licensing and authentication.

Technical improvements include:

- Attaining interoperability across DRM standards.
- Linking DRM standards across supply-chain wide functionalities.

The essential operative feature of this section is to create a fully interoperable mechanism for the proprietary and copyright aspects of digital content.

Section 8: Prototype Testing

This section involves building prototypes which work by application in the three inter-related software applications which the invention has been developing: CommonGroundPUBLISHER, CommonGroundLEARNER and CommonGroundCONFERENCE.

The essential operative feature of this section is to create a fully interoperable application which realises the potentials of the interlanguage apparatus in several specific areas of digital content management.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is now described in more detail in the form of non-limiting embodiments according to the present invention, clarified with the help of the enclosed drawings, where:

FIG. 1 illustrates the crosswalks dilemma, in which 17 ontologies require 272 crosswalks.

FIG. 2 illustrates the indefinitely extensible interlanguage mechanism—in which CGML is provided as an example.

FIG. 3 illustrates the interlanguage mechanism, by means of which the number of mappings equals the number of mapped schemas (in this case n =17)

FIG. 4 shows the method of operation of the Interlanguage apparatus.

FIG. 5 illustrates the ontology-building apparatus.

FIG. 6 illustrates one possible method of data entry within a publishing software system, the data from which can be exported into multiple ontologies using the underlying interlanguage invention.

FIG. 7 schematically illustrates one instance of abstract interlanguage representation, along with indicative tag synonyms.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

At the level of data, the interlanguage is a digital manufacturing mechanism. It is an invention which adds flexibility to the process of making a visible and represented meaning on a computer screen, a piece of paper or an audible sound. The manufacturing steps are as follows:

1. Data entry directly into an interlanguage interface, or into Ontology A, or and import of extant data created in Ontology A into the interlanguage;
2. Automated translation into interlanguage;
3. Translation from interlanguage into Ontology B;
4. Ontology B stylesheet creates a particular form of physical manifestation of communicated meaning for which Ontology B was designed, but not necessarily for which Ontology A was created.

At the level of metadata, the interlanguage is also a manufacturing mechanism, automatically allowing this metadata to be represented in a range of different ways by means of a data export apparatus as follows:

1. Metadata entry into an interlanguage interface, or into Ontology A, or and import of extant data created in Ontology A into the interlanguage;
2. Automated translation into interlanguage;

3. Export of data into Ontology B cataloguing, resource discovery or metadata database;
4. Rendering of metadata in formats characteristic of Ontology B, such as library cataloguing records or 'advanced' search mechanisms which are able to differentiate semantically different kinds of search.

One specific application of this invention is Common Ground Markup Language, an ontology of authorship and publishing which, by means of the interlanguage invention and the LEXICOGRAPHER apparatus, interoperates across seventeen major ontologies.

In the most challenging of cases—in which the raw digital material is created in a legacy DTD or ontology, and in which that DTD is not already known to the interlanguage from previous interactions—the invention:

i. interprets structure and semantics from the source DTD and its instantiation in the case of the particular quantum of source data, using the filter mechanisms described above—for example, in the case of publishing and the Common Ground Markup Language interlanguage, a hypothetical newly introduced digital rights management framework;

ii. draws inferences in relation to the digital rights DTD and the particular quantum of data, applying these automatically and presenting structured queries in cases where the apparatus and its filter mechanism 'knows' that supplementary human interpretation is required;

iii. stores any automated or human-supplied interpretations for future use, thus building knowledge and functional useability of this new DTD into the interlanguage—in this example, into Common Ground Markup Language. These inferences then become visible to subsequent users, and capable of amendment by users, through the CommonGroundLEXICOGRAPHER interface;

iv. interpellates the data into the interlanguage format, in this example Common Ground Markup Language;

v. creates a cross-walk from Common Ground Markup Language into a designated destination DTD, for instance a new format for structuring text for rendering to a flexible substrate, using the superordination and composition mechanisms—these are automated in cases where the structure and semantics of the destination DTD are self-evident to the apparatus, or they are the subject of structured queries where they are not, or they are drawn from the CommonGroundLEXICOGRAPHER memory in instances where the same query has been answered by an earlier user;

vi. Interpolates data into the destination format;

vii. Supplies data for destination uses—in this instance, digital rights data applied to a new rendering format.

To give a less challenging example, the source DTD can be already known to the interlanguage, by virtue of automated validations based not only on the inherent structure of the DTD, but also many validations against a range of data instantiations of that DTD, and also numerous user clarifications of queries. In this case, the source DTD might be the e-learning standard associated with the UK National Curriculum, and the destination DTD might be Educational Modelling Language.

In this case:

i. By entering data in an interface which 'knowingly' relates to an e-learning interlanguage, Learning Design Language, which has been created using the mechanisms of this invention, there is no need for the filter mechanisms nor the interpolation processes that are necessary in the case of legacy data and unknown source DTDs; rather data is entered directly into the interlanguage format, albeit through the user interface 'facade' of the source DTD—in this case, the UK National Curriculum Standard;

ii. the apparatus then interpolates the data onto the designated destination format, in this case, from the interlanguage of Learning Design Language, into Educational Modelling Language;

iii. The data can be used in the destination format, Educational Modelling Language.

iv.It is possible to use the interlanguage apparatus to construct and apply other meta-mark-up languages which tie together other semantically overlapping of contiguous ontologies. In each case, the invention construct in the interlanguage in part in automated ways, and in part by remembering and interpreting for later reapplication moments when a human response was required to a structured query. In this way the apparatus constructs an interlanguage appropriate to the particular range of required interoperabilities across a specified range of ontologies.

Another specific application of this invention is Learning Design Language, an ontology of curriculum documentation and pedagogy which, by means of the interlanguage invention and the LEXICOGRAPHER apparatus, interoperates across major e-learning and digital curriculum publishing ontologies.

It should be understood that the above description describes various embodiments of the invention. Clearly other variations which are understandable by a person skilled in the art without any inventiveness are included within the scope of this invention.

COMMON GROUND PUBLISHING PTY LTD
By its Attorneys
PIPERS (Melbourne)
Dated: 27 June, 2003

FIG. 1: The Crosswalks Dilemma Illustrated—17 Ontologies Require 272 Crosswalks

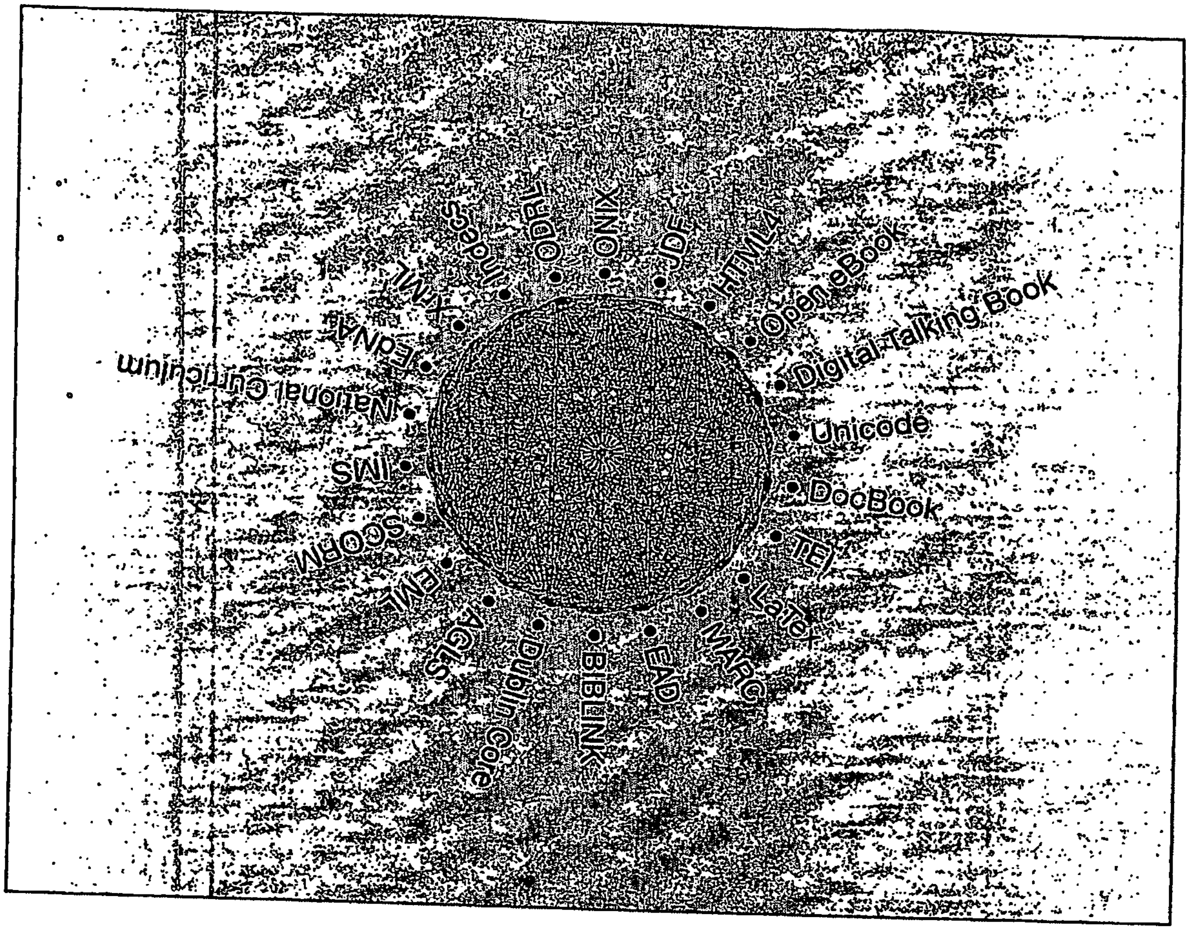

FIG. 2: The Indefinitely Extensible Interlanguage Mechanism—CGML as an example.

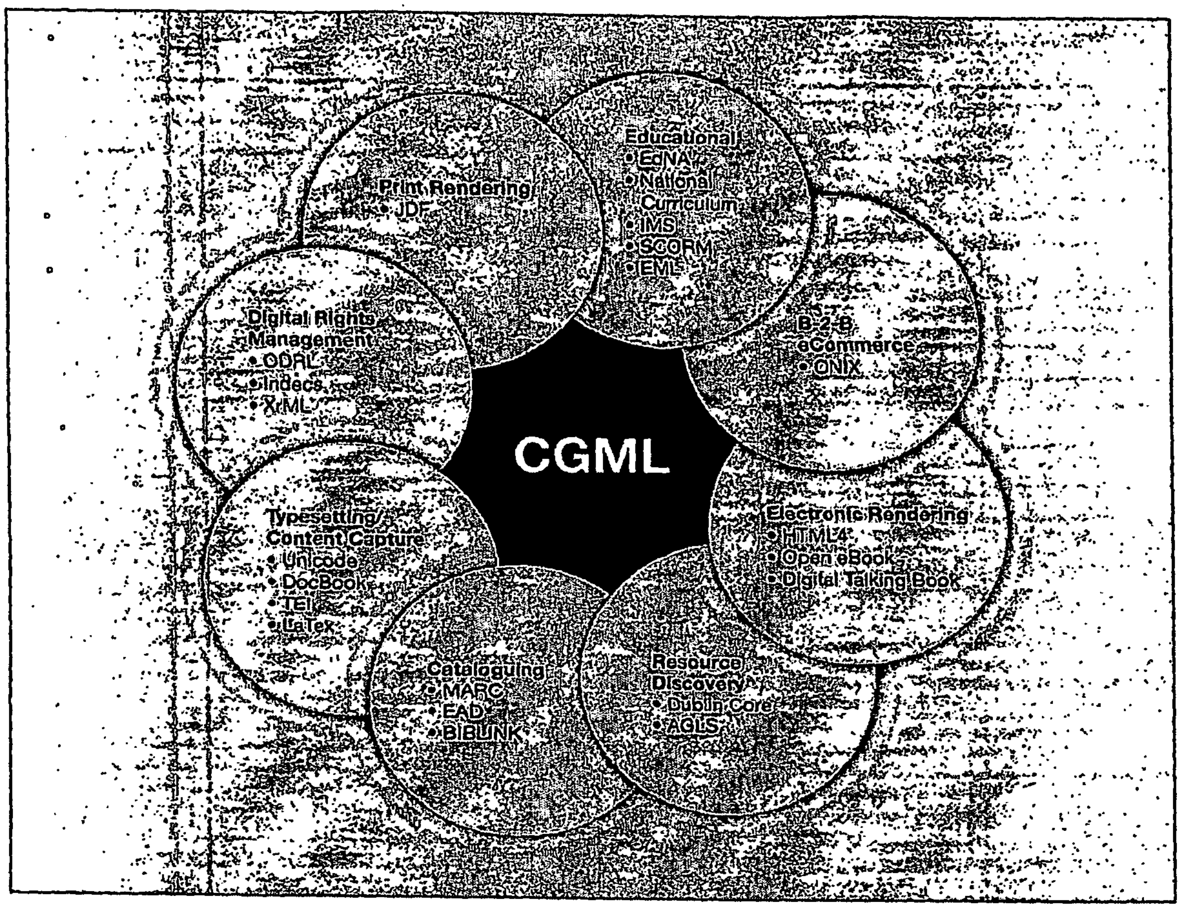

FIG. 3: Using the Interlanguage Mechanism, the Number of Mappings Equals the Number of Mapped Schemas (in this case n =17)

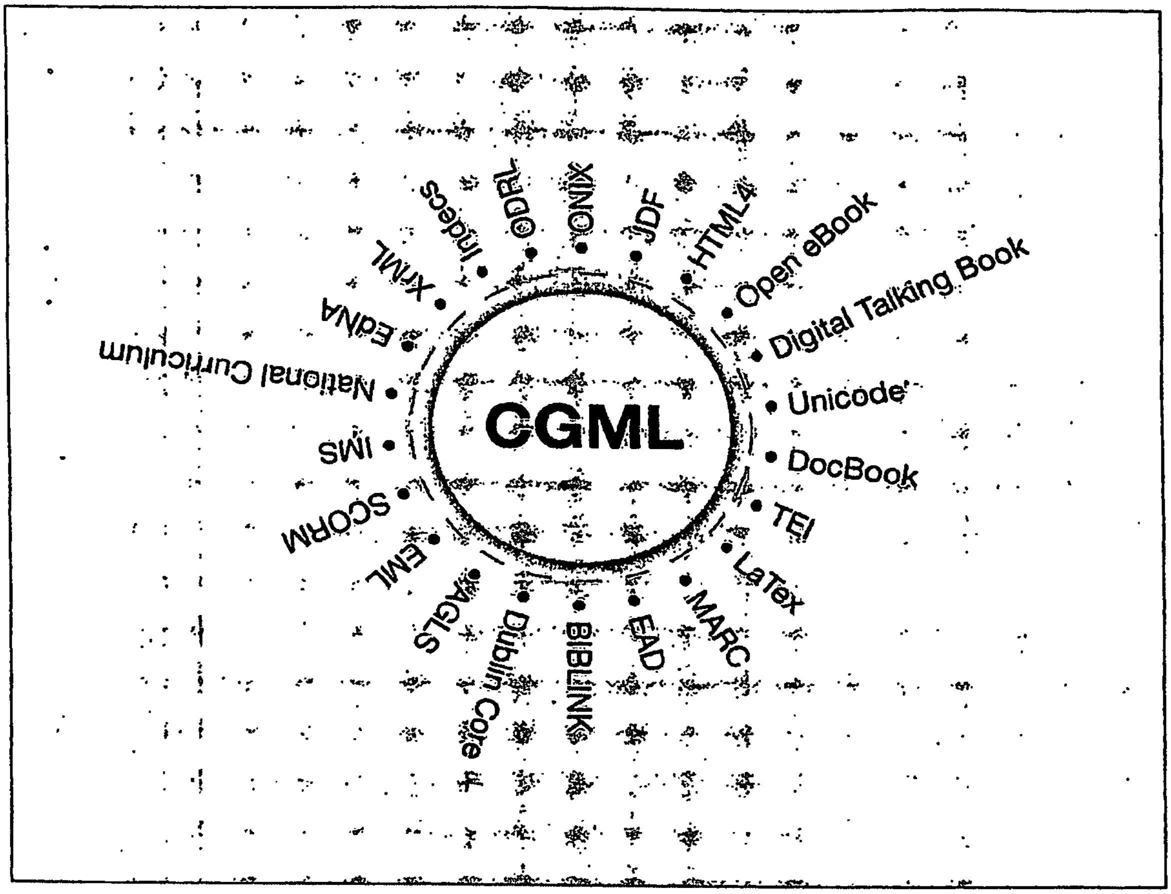

FIG. 4: The method of operation of the Interlanguage apparatus.

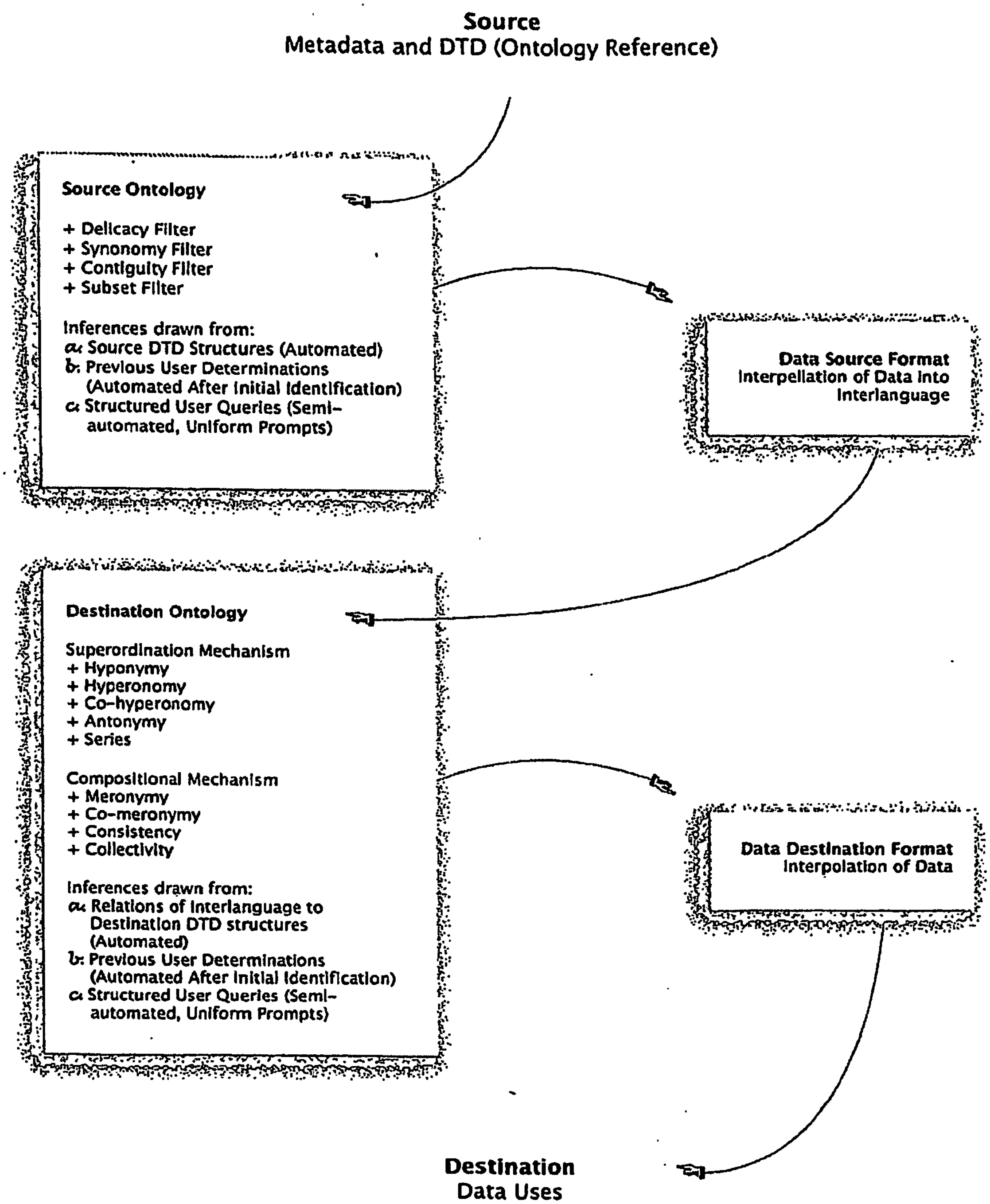

FIG. 5: Ontology-Building Apparatus—CommonGroundLEXICOGRAPHER.

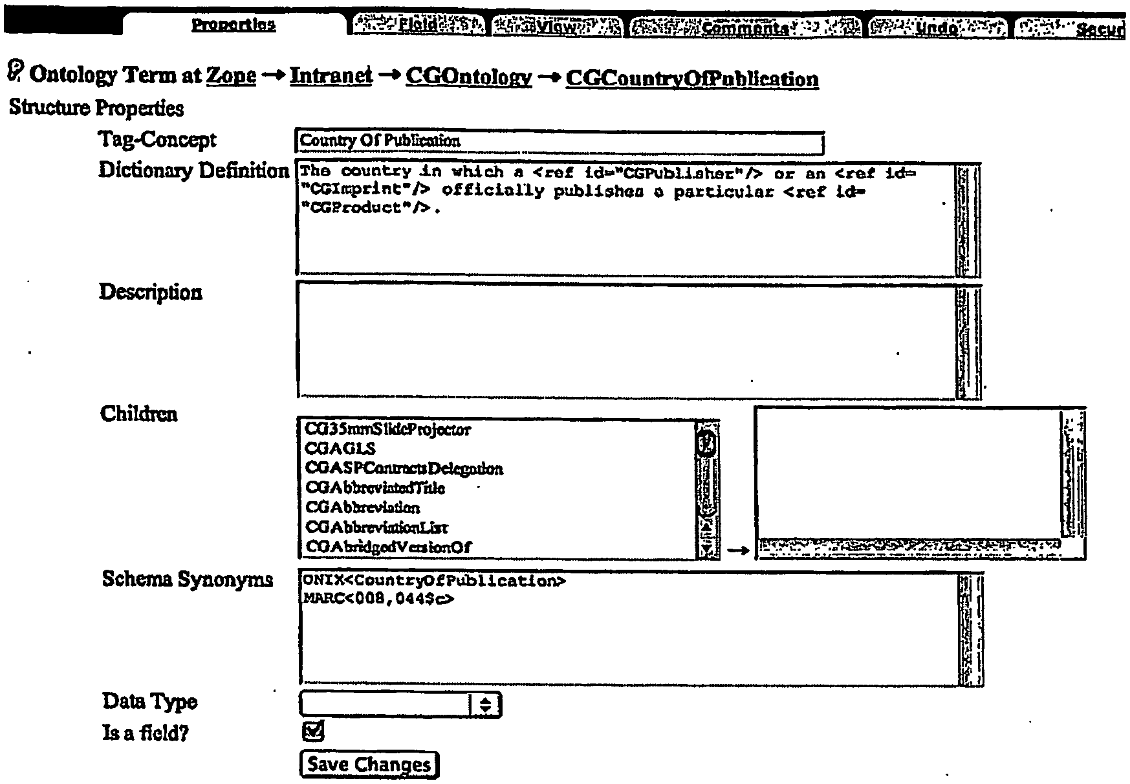

FIG. 6: Data entry area in CommonGroundPUBLISHER (formerly named the Creator-to-Consumer (C-2-C) System. Using the invention, data can be exported into multiple ontologies.

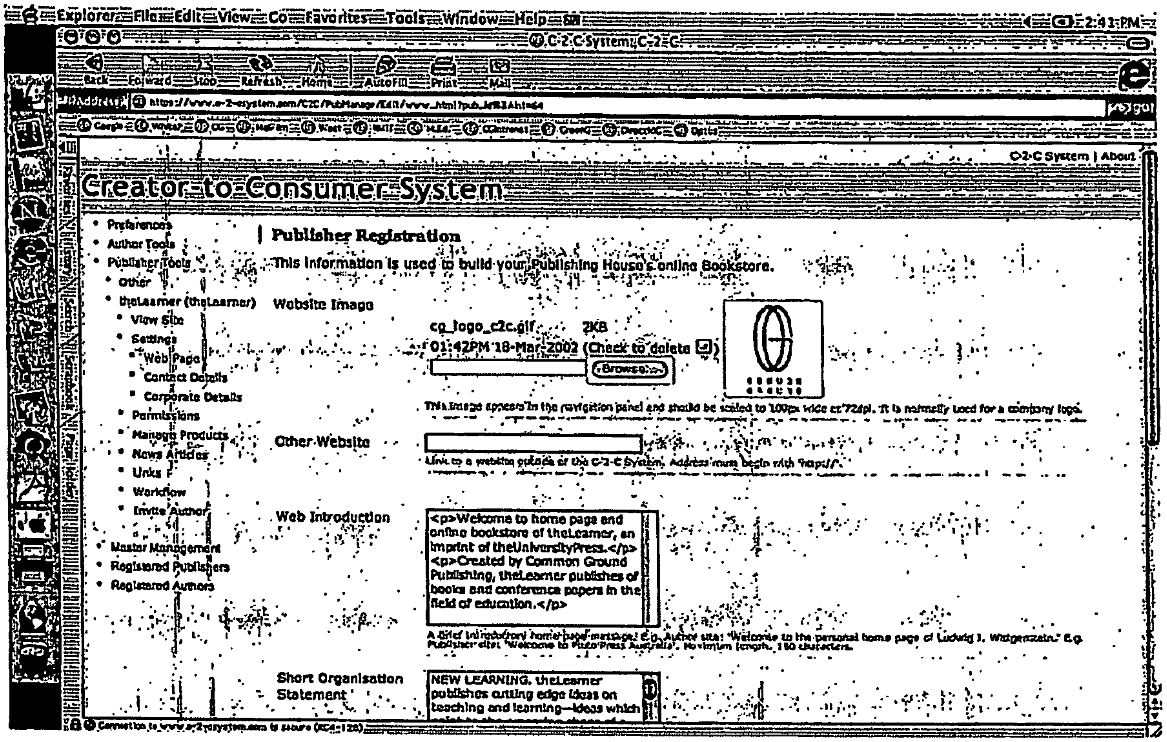

FIG. 7: Example of Taxonomic Representation of Tags Relations in overlapping Ontologies, as Generated by the Invention.

FOURTH ORDER CONCEPTS	FIFTH ORDER CONCEPTS	SIXTH ORDER CONCEPTS
CGParty XrML<principal> IN<party> SCORM<centity>		
	CGPerson IMS<person> ONIX<Name> XrML<commonName> IN<person> EAD<persname> MARC<100/700 I1 = 1; I2 = ø ; $a,600 I1 = 1; I2 = 4; $a >	
		CGNameFunction CGNameForPublication CGLegalName CGMailingName
		CGHonorific ONIX<TitlesBeforeNames> DB<honorific> MARC<100/700 I1 = 1; I2 = ø ; $a; $c,600 I1 = 1; I2 = 4; $c >
		CGGivenNames ONIX<NamesBeforeKey> DB<firstName>,DB<othername> MARC<100/700 I1 = 1; I2 = # ; $a,600 I1 = 1; I2 = 4; $a > EML<Initials-prefix>
		CGSurnamePrefix ONIX <PrefixToKey>
		CGSurname ONIX<KeyNames> DB<surname> EAD<famname> MARC<100/700 I1 = 1; I2 = # ; $a; $b,600 I1 = 1; I2 = 4; $a>
	V (to CGOrganisation)	CGNameSuffix ONIX<SuffixToKey> DB<lineage>